AD-A105 956   PRINCETON UNIV NJ DEPT OF STATISTICS   F/G 12/1
ASSESSING THE BEHAVIOR OF ROBUST ESTIMATES OF LOCATION IN SMALL--ETC(U)
MAR 81 D PREGIBON, J W TUKEY   DAAG29-79-C-0205
UNCLASSIFIED TR-182-SER-2   ARO-16669.6M   NL

END
DATE
FILMED
8
DTIC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 16669.6-M — 2. GOVT ACCESSION NO. AD-A105 956	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Assessing the Behavior of Robust Estimates of Location in Small Samples: Introduction to Configural Polysampling	5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Daryl Pregibon John W. Tukey	8. CONTRACT OR GRANT NUMBER(s) DAAG29 79 C 0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS Princeton University Princeton, NJ 08544	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709	12. REPORT DATE Mar 81 13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

D

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Configural polysampling offers an effective alternative to both usual Monte Carlo techniques and small sample asymptotics in studying and improving robust estimates Attention here is restricted to location estimators that are location and scale invariant. The methods are applicable to both simple and compound sampling situations. Several possible aspects of such studies include: (a) the determination of the minimum attainable variance in each sampling situation, (b) the determination of the maximum attainable polyefficiency over several sampling

20. ABSTRACT CONTINUED

situations, (c) the fine-tuning of robust estimates with the intent of increasing their polyefficiency for small n, and (d) the identification of data configurations where one can a priori expect poor performance with certain estimators.

6

Assessing the behavior of robust estimates of location in small samples: Introduction to configural polysampling*

by

10 Daryl Pregibon[redacted] John W. Tukey

15 DAAG29-79-C-0[illegible]

14 TR-182-SER-2

9 Technical Rep[redacted]t, No. 185, Series 2
Department of Statistics
Princeton University
11 Mar[redacted]81

12 16

18 ARO 19 16601.1-M

*Prepared in connection with research at Princeton University, supported by the Army Research Office (Durham).

**Present Address: Department of Biostatistics, University of Washington, Seattle 98195.

4[illegible]

ABSTRACT

Configural polysampling offers an effective alternative to both usual Monte Carlo techniques and small sample asymptotics in studying and improving robust estimates. We restrict our attention here to location estimators that are location and scale invariant. The methods are applicable to both simple (iid) and compound (non-iid) sampling situations. Several possible aspects of such studies include:

the determination of the minimum attainable variance in each sampling situation,

the determination of the maximum attainable polyefficiency over several sampling situations,

the fine-tuning of robust estimates with the intent of increasing their polyefficiency for small n,

the identification of data configurations where one can a priori expect poor performance with certain estimators.

Implementation of these ideas is being carried forward; results will be reported elsewhere.

1. Introduction.

A Robustness Week was held in Princeton in March 1980. Considerable emphasis was given to the question of what to do -- that is, which estimators to use -- in location problems for samples of sizes n=4 to 8 or 9. Two somewhat different schools of thought (and potential practice) emerged: (1) the theoretical study of estimators by small sample asymptotic methods, and (2) experimental study via configural-polysampling methods, to be discussed here. (The latter methods are by no means limited to such small samples.)

The methods we discuss are applicable to both simple and compound sampling situations. A simple situation consists of a collection of n independent random variables which are identically distributed:

$$X_i \sim F\left(\frac{x_i-\mu}{\sigma}\right) \qquad i=1,\ldots,n \ .$$

For example, the simple Gaussian situation consists of a random sample of size n from Gau(0,1). A sample of n from a mixture of Gaussian distributions would also correspond to a simple sampling situation. A compound situation consists of a collection of n independent random variables which are not identically distributed:

*Prepared in connection with research at Princeton University, supported by the Army Research Office (Durham).
**Present Address: Department of Biostatistics, University of Washington, Seattle 98195.

March 11, 1981

$$X_i \sim F_i\left(\frac{x_i-\mu}{\sigma}\right) \qquad i=1,\ldots,n \ .$$

For example, a realization from the one-wild Gaussian situation consists of a random sample of size n-1 from Gau(0,1) and a single Gau(0,100) variate. The use of compound situations is important in robustness studies though they have been used sparingly. For a recent account of some moment calculations for compound-situation order statistics, see Bruce, Pregibon and Tukey (1981).

For single situations (either simple or compound), estimator performance is usually measured by relative efficiency:

$$\mathrm{eff}_Q(t) = \frac{\text{min attainable variance in situation Q}}{\text{var(t|situation Q)}}$$

Configural sampling methods are useful in computing both the numerator and denominator of this quantity.

In terms of assessing the variability of estimators, configural sampling methods are expected to provide order-of-magnitude gains in precision, over usual Monte Carlo methods in the location problem (e.g. Hoaglin, 1971). Configural-polysampling methods also supply an honest estimate of the minimum attainable variance (subject to the usual invariance condition) for each sampling situation against which one can compare any desired estimator.

When several situations are considered, a useful measure of performance is the polyefficiency:

$$\text{polyeff}(t) = \min\{\text{eff}_A(t),\ \text{eff}_B(t),\ \ldots,\ \text{eff}_Z(t)\}$$

Configural-polysampling methods can be used to determine the maximum attainable polyefficiency.

More detailed data analysis of the configural-polysampling output for any particular estimator can suggest improvement to that estimator as a function of sample size (e.g. Should the tuning constant increase or decrease with n? Should the estimate be modified for small n?) and also as a function of particular configurations (For what sorts of configurations should we make what kinds of change?).

The opportunities for constructing confidence intervals (either conditional or unconditional) are also great. See (Tukey 1981a) for several possibilities.

2. Configural-sampling Methods.

2A. Conditional Mean Squared Error (MSE) Calculations.

Consider data $x_1,\ldots,x_n$ from situation Q, so that $X \sim \underline{F}_Q(\frac{x-\mu}{\sigma})$. Without loss of generality we will take $\mu=0$ and $\sigma^2=1$. Let $y_1 \le y_2 \le \ldots \le y_n$ denote the corresponding order statistics. The change of variables, for any $1 \le a \ne b \le n$:

$$r = y_a$$

$$s = y_b - y_a$$

$$c_i = (y_i - r)/s \quad \text{for } i\text{'s not } a \text{ or } b\ .$$

has Jacobian s^{n-2}. The vector c will be called the sample configuration (this is one of many choices for a definition of a location-and-scale configuration). As we are concerned only with location-and-scale invariant estimators, we have

$$t = t(y) = t(r+sc) = r + s\cdot t(c) \ .$$

Thus we may think of our estimator t as operating on the configuration vector since, conditional on the observed configuration, t(c) is constant. This implies that the conditional MSE{t(y)} is given by

$$MSE\{t(y)|c\} = E_{r,s}\{r + st(c)|c\}^2 \ . \tag{1}$$

This is a quadratic function of t(c) and is minimized by

$$t_o(c) = E\{rs|c\}/E\{s^2|c\} \tag{2}$$

leading to the optimally invariant estimate of μ given by

$$t_o(y) = r + st_o(c) \ .$$

This estimate has conditional MSE given by

$$MSE\{t_o(y)|c\} = E\{rs|c\}t_o(c)+E\{r^2|c\} \ . \tag{3}$$

Note that each of the above conditional expectations are naturally evaluated as two dimensional integrals and can be computed numerically with high precision (see Hoaglin, 1971, and Relles and Rogers, 1973, for the case a=1, b=n).

2B. Unconditional Monte Carlo Estimation.

For a specific configuration, the conditional mean squared error of t_o is given by (3). The estimate $t_o(y)$ is not generally conditionally unbiased for μ, though the unconditional bias vanishes for symmetric situations. Thus, the unconditional minimum attainable variance at situation Q is given by

$$\mathrm{var}\{t_o(y)\} = E_c\{\mathrm{MSE}\{t_o(y)|c\}\} = E_c E_{r,s|c}\{t_o^2(y)|c\} \quad .(4)$$

This value can be approximated by averaging $\mathrm{MSE}\{t_o(y)|c\}$ over a sample of configurations drawn from situation Q, leading to the sampling estimate

$$\hat{V}\mathrm{ar}\{t_o(y)\} = \underset{c}{\mathrm{ave}}\ \mathrm{MSE}\{t_o|c\} \ . \qquad (5)$$

2C. Assessment of other Estimators at each Situation.

A location and scale invariant estimator $t(y)$ is completely determined by specifying the value $t(c)$ associated with a particular configuration. The conditional mean-squared-error of the resulting estimate $t(y) = r + st(c)$ can be determined from (1), which simplifies into one quadratic function for each situation, (whose choice affects MSE, E and t_o):

$$\mathrm{MSE}\{t(y)|c\} = \mathrm{MSE}\{t_o(y)|c\} + E\{s^2|c\}(t(c)-t_o(c))^2 \quad .(6)$$

The overall assessment of $t(y)$ at situation Q is obtained by averaging (6) over configurations:

$$\hat{v}ar\{t(y)\} = \underset{c}{ave}\ MSE\{t(y)|c\} .$$

This value is to be compared to the Monte-Carlo estimate of the minimum attainable variance given by (5). This leads to the measure of invariant Q-efficiency

$$\hat{e}ff_Q\{t(y)\} = \hat{v}ar\{t_o(y)\}/\hat{v}ar\{t(y)\} ,$$

where Q refers to the particular situation at hand (for example either G = Gaussian, S = slash, or W = one-wild). This estimate of the Q-efficiency differs from the more usual bounds on Q-efficiencies offered by (a) an estimated variance and the Cramer-Rao lower bound (which is exact for small samples only if some estimate $t_o(y)$ attains the lower bound), or (b) the upper bound (relative Q-efficiency) based on the estimated variance and the estimated best of the variances at the situation among estimates so far evaluated. The calculations of Hoaglin (1971) suggest that for small n and heavy tails, the conservativeness of the Cramer-Rao bound can be considerable - too small by a factor of 2.

3. Configural-polysampling Methods.

So far, we have treated a specific situation: (a) determining the best one can do and (b) assessing the performance of a non-optimal estimator at that situation, both absolutely and compared to the best. Often one is, or should be, interested in the performance of estimators at several distinct situations. Conventionally, samples are drawn from each situation, and some measure of performance is computed at each. Such studies can be made

1) more efficient, and

2) more informative

by employing configural polysampling methods. As the name implies, configurations rather than samples are used throughout. In order to use a common collection of configurations, one employs configurations some of which are obtained from situations other than the particular situation being studied. This is possible only if proper weights are attached to each configuration, for use at each situation. (These weights are then functions of both situation and configuration.) In this section, we first show the relevance of these methods in assessing single-situation performance. We then discuss their importance in assessing polysituation performance.

3A. Single-situation Performance.

Consider a single situation A, $\text{eff}_A(t)$ being the measure of performance. If all samples are drawn from situation A, then the methods of section 2 provide the configural sampling estimate of $\text{var}_A(t)$ given by

$$\hat{\text{var}}_A(t) = \underset{c}{\text{ave}}\ \text{MSE}_A\{t(y)|c\} = \int_c \text{MSE}_A\{t(y)|c\}dF_N(c)$$

where $dF_N(c)$ is the empirical distribution function based upon N realizations of c. Now consider the implications of estimating $\mathrm{var}_A(t)$ when sampling from situations other than A, say situation Q. If we put

$$W_{A|Q} = dF_A(c)/dF_Q(c) = f_A(c)/f_Q(c)$$

the method of importance sampling (e.g. Hammersley and Handscomb, 1970) provides the expression

$$\mathrm{var}_A(t) = \mathrm{var}_{A|Q}(t) = \int_c \mathrm{MSE}_A\{t(y)|c\} W_{A|Q} dF_Q(c)$$

which is estimated by the weighted average over configurations

$$\hat{\mathrm{v}}\mathrm{ar}_A(t) = \hat{\mathrm{v}}\mathrm{ar}_{A|Q}(t) = \underset{c}{\mathrm{wave}}\ \mathrm{MSE}_A\{t(y)|c\}$$

where the weights implied in "wave" are, of course, $W_{A|Q}$. In general, we want to consider polysampling at situations A through Z, drawing N_A through N_Z configurations totalling $N = \sum_Q N_Q$, with fractions $\alpha_Q = N_Q/N$. The variance of t satisfies

$$\mathrm{var}_A(t) = \sum_Q \alpha_Q \mathrm{var}_{A|Q}(t)$$

and can, in particular, be estimated by

$$\hat{\mathrm{v}}\mathrm{ar}_A(t) = \sum_Q \alpha_Q \hat{\mathrm{v}}\mathrm{ar}_{A|Q}(t)$$

Note that $W_{A|Q}$ is not necessarily bounded for given A and Q -- and is likely to be unbounded. Thus, in extreme configurations, where the likelihood ratio f_A/f_Q is very large, the contribution to the estimated variance will be disproportionately large. In such cases, the variance reduction properties of importance sampling are not maintained and the technique could be terribly inefficient.

In order to develop a more useful set of weights, regard the $N_A+N_b+\ldots+N_Z$ configurations as a restricted (by quotas for A,B,...,Z) sample of configurations from

$$H(c) = \alpha_A F_A(c) + \ldots + \alpha_Z F_Z(c)$$

where the α_Q's with $\Sigma\alpha_Q=1$, are as above. Then, for evaluation at situation A and sampling from H(c), we can put

$$W_{A|H} = dF_A(c)/dH(c) = f_A(c)/\Sigma\alpha_Q f_Q(c) \quad .$$

and write

$$\begin{aligned} \mathrm{var}_A(t) &= \int_c \mathrm{MSE}_A\{t(y)|c\}dF_A(c) \\ &= \int_c \mathrm{MSE}_A\{t(y)|c\}W_{A|H}dH(c) \end{aligned}$$

This expression can be rearranged to give

$$\begin{aligned} \mathrm{var}_A(t) &= \Sigma\alpha_Q\int_c \mathrm{MSE}_A\{t(y)|c)\}W_{A|H}dF_Q(c) \\ &= \Sigma\alpha_Q \ \mathrm{var}^*_{A|Q}(t) \end{aligned}$$

which can be estimated by

$$\hat{v}\mathrm{ar}_A(t) = \Sigma\alpha_Q\hat{v}\mathrm{ar}^*_{A|Q}(t)$$

where

$$\hat{\mathrm{V}}\mathrm{ar}^{*}_{A|\Omega} = \underset{c}{\mathrm{wave}}\ \mathrm{MSE}_{A}\{t(y)|c\}$$

with weight $W_{A|H}$. Thus we obtain an estimate of form similar to our previous one, though now the weights are bounded. This choice of weights will provide stable contributions to the overall variance estimate even in cases where extreme configurations are observed.

3B. Polysituation Performance.

In polysampling studies one is primarily interested in assessing the performance of an estimator over all the small collection of situations considered. Naturally, to judge the usefulness of an estimator, one must know how well the best estimator performs across these situations. For example, for n=20 and the Gaussian, slash and one-wild-Gaussian situations, the best known polyperformance (as measured by relative tri-efficiency) is about 93%. This does not correspond to the optimal estimate for each situation, only the best estimate known to date.

Configural-polysampling can help remove the arbitrariness of the "best known" attainable performance, replacing it by the "best" attainable performance (to the accuracy provided by sampling). When polyperformance is a function of the various individual-situation sampling variances, Tukey (1981b) describes a possible iteration within the family of Sayeb estimates. Essentially the method requires finding the "best" estimate for each configuration, across situations. This estimate is not determined by the behavior at this configuration alone; iterative refinement on its values based on behavior at all configurations is required.

As configural polysampling determines overall estimator performance as a combination of results based on individual configurations, the possibility of modifying existing estimators to obtain improved performance is an attractive feature. A tentative approach would rely on some measure(s) which summarize configuration structure (see, for example, Tukey 1981b) and then attempt to correlate these with the estimator-performance results. Of particular interest is the determination of configuration types where the estimator performs poorly. The usefulness of a simple-configuration summary characteristic may lead to new types of adaptive estimates, or adaptive modifications to existing estimates. See for example Bell's (1980) work in adaptive biweight scale estimation.

REFERENCES

Bell, B. (1980). "Adaptive scale estimation in the location problem," Technical Report, Stanford University.

Bruce, A., Pregibon, D. and Tukey, J. W. (1981). "The second representing function for compound situations, Technical Report No. 186, Series 2, Department of Statistics, Princeton University.

Hammersley, A. and Handscomb, A. (1971). Monte Carlo Methods, Oxford University Press, Oxford, England.

Hoaglin, D. (1971). "Optimally invariant estimators of location," Unpublished Ph.D. thesis, Princeton University.

Relles, D. and Rogers, W. (1973). "Are statisticians robust estimators of location," JASA.

Tukey, J. W. (1981a). "Kinds of polyconfidence limits for centers, and some thoughts on identification and selection of confidence procedures using polysampling," Technical report No. 190, Series 2 Department of Statistics, Princeton University.

Tukey, J. W. (1981b). "Some advance thoughts on the data analysis involved in configural-polysampling directed toward high-performance estimates," Technical Report No. 189, Series 2, Department of Statistics, Princeton University.

Tukey, J. W. (1980). "Speeding up configural polysampling evaluation," unpublished memorandum, Department of Statistics, Princeton University.

March 11, 1981